Five Urban Legends of CNS Clinical Trial Methodology: Unsuccessful Solutions to the Problem of Failed Trials

INTRODUCTION
The problem of failed trials in CNS is well recognized (Khin et al. 2011). As the number of failed trials has grown, drug developers have attempted several strategies to improve signal detection and reduce the failure rate. We present five common strategies and evaluate the evidence for their effectiveness.

URBAN LEGEND #1 INCREASING SAMPLE SIZE WILL INCREASE STATISTICAL POWER

If one assumes statistical power increases with sample size and effect size is fixed, it appears reasonable to assume that increasing sample size will increase effect size in a given study.

Testing the legend:
Liu et al. (2008) examined four randomized, double-blind placebo controlled phase III depression trials with at least 150 subjects per treatment arm to investigate the effect of cumulative sample size on drug-placebo difference.

The results:
Site performance across consecutive studies of fluoxetine was inconsistent, with <0.50 correlation within sites across studies on randomization rates, protocol completion percentage, percentage of placebo responders and drug-placebo differences.

Summary:
Since most sites only enroll 5-10 subjects per study arm, they are underpowered to demonstrate drug-placebo differences. However, many sites experience considerable personnel turnover from year to year, which may affect their success rate. Good performance of a site in one clinical trial had a very low correlation with success rate in other trials, which may affect their success rate.

URBAN LEGEND #2 CHOOSING THE “RIGHT” SITES WILL REDUCE RISK OF FAILURE

There is a belief that selecting investigative sites with proven results will continue to provide positive results.

Testing the legend:
Gelhock et al. (2002) analyzed data from 21 clinical trial sites that participated in at least two different trials and randomized at least 30 subjects across trials.

The results:
Site performance across consecutive studies of fluoxetine was inconsistent, with <0.50 correlation within sites across studies on randomization rates, protocol completion percentage, percentage of placebo responders and drug-placebo differences.

Figure 2: Correlations of the same metrics for the same sites in consecutive trials

URBAN LEGEND #3 USING THE MOST EXPERIENCED RATERS WILL REDUCE RISK OF FAILURE

It seems logical that employing more experienced raters will minimize variability and improve signal detection.

Testing the legend:
Kobak et al. (2006) examined the relative impact of experience and calibration by calculating interrater agreement across three groups of raters: an experienced and calibrated cohort, an experienced but non-calibrated cohort, and an inexperienced cohort. Thirty subjects with MDD were assessed in independent interviews by two different raters on the same day using the Structured Interview Guide for the Hamilton Depression Scale (SIGH-D).

The results:
The highest interrater agreement was achieved by experienced and calibrated raters (>0.93), followed by inexperienced raters (>0.77). Experienced but non-calibrated raters achieved the lowest interrater agreement (>0.55).

Figure 3: Intraclass correlation coefficient (ICC) values of interrater agreement of three rater cohorts

Summary:
Experience alone does not result in good interrater reliability within a cohort of raters. Experienced raters must be carefully calibrated with each other to achieve this.

URBAN LEGEND #4 INCREASING RATER TRAINING WILL REDUCE RISK OF FAILURE

A frequently cited cause of trial failure is inadequate rater training. There is the potential for huge variability across raters in a single trial, which negatively affects study power and signal detection. Some believe increasing the intensity of rater training will reduce variability.

Testing the legend:
Dorrestein et al. (1998) trained 85 raters on the HAM-A in an intensive, six hour iterative training session with four videotapes and discussion between each tape. Training consisted of a lecture on the HAM-A, a detailed review of each individual item and how to rate it, and a review of a training manuscript designed for the particular study.

The results:
ICC across the four training tapes ranged from 0.65-0.79 and did not improve across the six hours of reliability training.

Figure 4: Relationship between baseline rating and treatment effect (HAMD score and Clinical Global Impression - Improvement [CGI-I])

URBAN LEGEND #5 CERTAIN REGIONS OF THE WORLD HAVE BETTER SIGNAL DETECTION

Many researchers believe that greater signal detection can be obtained outside the US.

Testing the legend:
Khin et al. (2011) conducted a meta-analysis of 81 randomized double-blind clinical trials of antidepressants that were submitted to the FDA between 1983 and 2008, including both US and ex-US studies. Another meta-analysis (Pohn et al. 2005) looked at 33 randomized double-blind clinical trials of schizophrenia.

The results:
Both meta-analyses documented increasing placebo response across both US and ex-US regions, and a decrease in treatment effect for US studies.

Figure 5: Placebo response and treatment effects over time in US and non-US MDD trials.

Summary:
Although there is regional variability, the failure rate of studies of depression and schizophrenia outside the US is increasing.

CONCLUSIONS

Strategies for improving signal detection in CNS clinical trials are often used without clear evidence of their efficacy. Increasing sample sizes, targeted site selection, using experienced but non-calibrated raters, increasing rater training, and conducting trials outside of the US have not proven successful in increasing the success rate. These “urban legends” are widely touted, but evidence to support them is lacking.

References:

Disclosures: